Detection Model for URL Phishing with Comparison Between Shallow Machine Learning and Deep Learning Models

Nizam Aditya Zuhayr[1(✉)], Girinoto[2], Nurul Qomariasih[2], and Hermawan Setiawan[2]

[1] The Center for Research and Development of Cyber and Crypto Security Technology, National Cyber and Crypto Agency Jakarta, Jakarta, Indonesia
nizam.aditya.1@gmail.com

[2] Crypto Software Engineering, State Cyber and Crypto Polytechnic Bogor, Bogor, Indonesia
{girinoto,nurul.qomariasih,hermawan.setiawan}@poltekssn.ac.id

Abstract. In the report on trends in phishing activity released by the Anti-Phishing Working Group (APWG), global phishing cases continued to increase throughout 2021 to the first quarter of 2022. This study compares shallow machine learning algorithms that have been used by governments with deep learning in classifying URLs. Phishing. From the data as many as 30,047 URLs consisting of 15,022 phishing URLs and 15,025 legal URLs, the distribution was carried out for training data and test data. URL phishing modeling uses deep learning algorithms LSTM and GRU as well as the best shallow machine learning algorithms from research conducted by Rao et.al, namely Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT). Modeling is done based on URL characteristics, text structure, and a combination of URL characteristics with text structure. Based on URL characteristics, the model with the best accuracy from the shallow machine learning algorithm is Random Forest at 97.4%, while the deep learning algorithm is LSTM at 96.7%. Based on the structure of the text, the best deep learning algorithm is the GRU of 97.8%. While the combination model using 2 deep learning algorithms LSTM and GRU get an accuracy of 98.1%. Furthermore, the combination model as the best model is implemented in the form of a website using the Flask framework with the classification results in the form of a URL probability score that is detected as a phishing URL.

Keywords: phishing · machine learning · deep learning · classification model · flask

1 Introduction

The COVID-19 pandemic has forced the implementation of a new life adaptation (IMR) and urged people to work from home. In this situation, information and communication technology has an important role in meeting daily needs [1]. The number of phishing attacks has been steadily increasing since the COVID-19 outbreak in late 2019 [2]. Phishing is a way to steal someone's credentials. In the phishing activity trend report

I. H. Agustin (Ed.): ICONNSMAL 2022, AISR 177, pp. 146–156, 2023.
https://doi.org/10.2991/978-94-6463-174-6_13

issued by the Anti-Phishing Working Group (APWG), global phishing cases continued to increase from 2021 to the first quarter of 2022 [3]. In the first quarter of 2022, 1,025,968 phishing attacks were recorded, an increase from 888,585 attacks in the 4th quarter of 2021.

Phishing is a social engineering technique to steal user credentials [4]. Perpetrators who engage in such acts usually use different techniques and psychological factors to convince victims to click on phishing links [5]. When opening a phishing link or phishing website, the system prompts the user to enter their credentials [2].

24,298 URLs consisting of 15,022 phishing URLs and 9,276 legal URLs. There is a difference of 5,746 data between phishing data and not. In order for the data to be balanced, legal URL data is added from open sources [6].

In reducing and overcoming the negative impact of phishing URLs (Uniform Resource Locator), the government, in this case, the Indonesian National Cyber and Crypto Agency through one of its parts, namely the Center for Research and Development of Cyber and Crypto Security Technology, conducted a study on phishing URL detection. The methods used are shallow machine learning algorithms such as Logistic Regression and Multinomial Naive Bayes.

Machine learning is a sub-field in computer science that is related to creating algorithms for specific purposes that depend on data sets [7]. In machine learning, there are 2 approaches, shallow learning (supervised, unsupervised, reinforcement) and deep learning [7]. Deep learning is defined as a learning method carried out by machines by imitating the basic working system of the human brain or commonly called a neural network using modern algorithms and mathematical tools separately [7]. Deep learning is one such approach that has better results than the shallow machine learning approach if large amounts of data are available [8].

Phishing URL detection can be done using various methods, namely based on the website URL text, HTML content, behavior, or a combination of the three methods [8, 9]. Detection of phishing URLs based on URLs can be done using 2 approaches, namely based on lexical features and the structure of the URL text itself.

Research conducted by Rao et al. using shallow machine learning shows that the Random Forest, Logistic Regression, and Decision Tree algorithms are the best algorithms. This method detects phishing URLs with an accuracy of 94%, 92%, and 91% respectively [6]. Research conducted by Su Yang shows that LSTM produces learning with an accuracy of 99% [10]. This research proposes the creation of deep learning and shallow machine learning models from datasets owned by the Center for the Study and Development of Cyber and Password Security Technology as a solution to problems in detecting phishing URLs. Of the algorithms used, the best algorithm is selected to be implemented in detecting phishing URLs.

2 Related Research

Su's research detects phishing URLs using LSTM. The advantage of using LSTM is that it has long-term dependencies. LSTM can learn data characterization automatically without manual complex feature extraction and has strong potential in dealing with complex massive data. Experimental results show that this model is close to 99.1% accuracy, higher than other neural network algorithms.

Rao's research detects phishing URLs using 6 shallow machine-learning algorithms. These algorithms are XGBoost, Random Forest (RF), Logistic Regression (LR), K-Nearest Neighborhood (KNN), Support Vector Machine (SVM), and Decision Tree (DT). The algorithm with the best result is Random Forest with 94.25% accuracy. The dataset is obtained from crawling on common sites and Alexa for non-phishing URLs and phishing URLs from PhishTank [6].

Korkmaz's research made comparisons for phishing URL detection using 8 shallow machine-learning algorithms. These algorithms are Logistic Regression (LR), K-Nearest Neighborhood (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), XGBoost, Random Forest (RF), and Artificial Neural Network (ANN) [11].

This study also uses 3 different datasets to compare each shallow machine learning algorithm. The first dataset is non-phishing URLs obtained from the Alexa database and phishing URLs from PhishTank. The second dataset is non-phishing URLs obtained from crawling on common sites and phishing URLs from PhishTank. The third dataset is non-phishing URLs from crawling on common sites and Alexa while phishing URLs from PhishTank [12].

This research was conducted by Ozcan et al. in 2021. This research makes a comparison between shallow machine learning and deep learning. Deep learning gets superior results in accuracy. The deep learning approach used is to combine natural language processing and vectorization feature extraction techniques as input to deep learning models. The best accuracy performance is obtained by the Deep Neural Network and LSTM architectures. The result after doing Hyperparameter tuning is to get 99% accuracy.

3 Research Methodology

The research data used in this study is divided into two sources. The first data was obtained from the BSSN the Center for the Study and Development of Cyber and Password Security Technology locus. The second data uses open-source data [6]. The data is in the form of legal URL samples and phishing URLs. The data from the two sources obtained are processed to become one.

This stage makes adjustments to the data so that it can be processed by deep learning. Adjustments are made by converting text data into integers. Phishing URLs have certain anomaly features that distinguish them from others that are not. Korkmaz et al. use features such as domain length, number of slashes in URLs, number of special characters, and so on to detect anomalies in URLs [11]. Feature extraction is performed on the data as an indication of an anomaly. The data is divided into training data and test data. The training data is used for modeling deep learning and shallow machine learning algorithms. The test data is used to evaluate the model that has been trained.

Modeling is used to conduct training on training data so that the results are in the form of a model that can detect phishing URLs. The models to be used are the RNN, LSTM, and GRU deep learning algorithms. For comparison, the 3 best shallow machine learning models were used from research conducted by Rao et.al using Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT) algorithms [6] (Fig. 1).

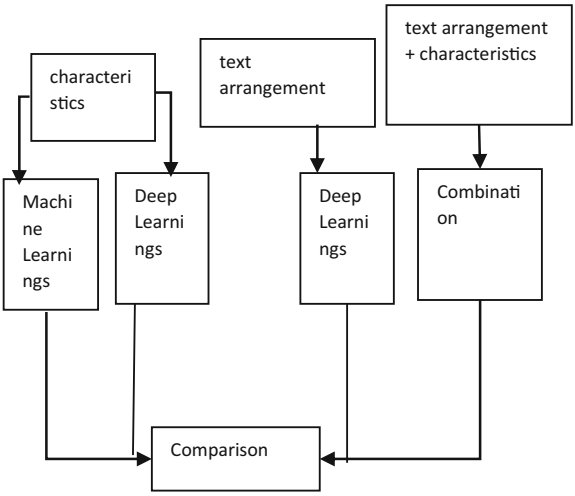

Fig. 1. Model comparison flow chart

This stage tests the prediction model that has been made before. This test is carried out using test data. At this stage, hyperparameter tuning is also carried out to get the best detection model.

The existing detection model is then evaluated using the confusion matrix to obtain accuracy, precision, and recall as well as the AUC score. These results are used to make comparisons of 3 shallow machine learning models, namely RF, LR, DT, and 3 deep learning models, namely RNN, LSTM, and GRU.

4 Results and Discussion

DBased on research [6, 11], and [13] there are 18 features that can be obtained from a URL. Based on these 18 features, categories c to r can be extracted from each URL structure, such as full URL, domain, subdomain, and path. So that the features that can be obtained are 66 features. There are 2 additional features from [11] that can be used based on 66 features. The data type consists of objects and integers, then the data type is changed from object to integer so that it can be processed by deep learning. The changed features are protocol and TLD (Top Level Domain). The Protocol is changed using one-hot-encoding, while TLD is changed using label-encoder. The result of one-hot-encoding adds features to 69 features.

Based on the URL text, the data used is the arrangement of the alphabet and characters in the URL. URLs in the form of strings will be converted into numbers using vectorization techniques. The technique used is the bag of words which consists of 3 stages namely, tokenizing, vocabulary creation, and vector creation [14]. In this research, the token is used in the letter fragments in the URL.

Based on the characteristics of the URL, the created model is used to predict phishing or legal URLs based on the features obtained from the URL. There are 69 features used,

based on the results of the preprocessing data. The classification model was created using the RNN, LSTM, and GRU architectures.

The model is used to carry out the phishing URL classification process or not. Modeling in this study was carried out using the RNN, LSTM, and GRU architectures. In addition, a shallow machine learning algorithm will also be used using the Random Forest, Logistic Regression, and Decision Tree algorithms as a comparison [15]. The results of the modeling will get accuracy, and AUC scores to compare the results of the analysis. Each algorithm will be hyper-parameter tuning to get the results.

4.1 Based on URL Characteristics

The Evaluation was carried out on random forest shallow machine learning algorithms, Logistic Regression, and Decision Tree as well as RNN, LSTM, and GRU deep learning algorithms.

The results of the classification using 3 shallow machine learning algorithms are in Table 1.

In optimizing the previous model, hyperparameter tuning is done. This study utilizes the grid search library to find the best hyperparameters. In the Random Forest algorithm, the hyperparameters that are tuned are 'max_features' and 'n_estimators', in Logistic Regression, the hyperparameters that are tuned are 'solvers', 'penalty', 'C value', while in the Decision Tree, the tuning is done on 'criterion ', and 'max_depth'.

From Table 2 it can be seen that there has been an increase in scores for all aspects of the assessment after hyperparameter tuning, starting from accuracy, precision, recall,

Table 1. Benchmarking table shallow machine learning before hyperparameter tuning

	Random forest	*Logistics Regression*	*decision tree*
Accuracy	**0.971**	0.927	0.958
Precision	**0.971**	0910	0.959
recall	**0.971**	0947	0.951
F1 scores	**0.971**	0.928	0.957
AUC scores	**0.995**	0.971	0.957

Table 2. Shallow machine learning Benchmarking table after hyperparameter tuning

	Random forest	*Logistics Regression*	*decision tree*
accuracy	**0.974**	0.928	0.958
Precision	**0.971**	0912	0.960
recall	**0.976**	0.946	0.954
F1 scores	**0.974**	0.929	0.957
AUC scores	**0.996**	0.971	0.973

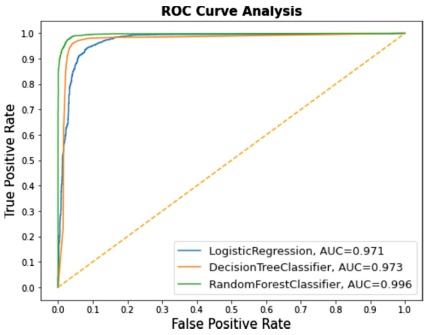

Fig. 2. ROC Curve Analysis.

and f1 score, to the AUC score [16]. The Random Forest algorithm remains the best algorithm for classifying phishing URLs. A Comparison of the ROC curve can be seen in Fig. 2.

The ROC curve in Fig. 2 shows that the closer the line is to 1, the better the model is at making predictions. The blue curve line (Logistic Regression) gets the lowest result, and the green curve line (Random Forest) gets the highest result.

The input to the first layer in the RNN architecture consists of 69 features which are parameters in the URL, then enter into the RNN layer [17]. After leaving the RNN layer then enter the dropout layer. In the dropout layer, disposal is carried out from the unit randomly. This is to prevent the model from overfitting. Overfitting is a condition where the model is too good to classify the data. After the dropout layer, enter the RNN layer and the output will be 32 units. The results will enter the dense layer with 32 units and finally 1 unit for classification.

The input on the first layer consists of 69 features, then it enters the LSTM layer [18]. After leaving the LSTM layer then enter the dropout layer and enter the Bidirectional LSTM layer. This layer is a development of LSTM, the difference is that the Bidirectional LSTM layer has 2 models at once. The first model learns the input sequence given, and the second model learns the reverse of the input sequence. After that go to the dropout layer and enter the Bidirectional LSTM layer again and enter the dense layer.

The results of the classification using the deep learning algorithm are in Table 3.

The results of testing the 6 methods obtained results which can be seen in Table 4.

Table 3. Deep Learning Methods

	RNN	LSTM	GRU
accuracy	0.938	**0968**	0.959
Precision	0.929	**0.963**	0.956
recall	0.949	**0.973**	0962
F1 scores	0939	**0968**	0.959
AUC scores	0978	**0.994**	0989

Table 4. Compare 6 methods Random Forest (I), Logistic Regression (II), Decision Tree (III), RNN(IV), LSTM(V), and GRU (VI) [20]

	I	II	III	IV	V	VI
accuracy	**0.974**	0.929	0.958	0.938	0968	0.959
Precision	**0.971**	0.919	0.954	0.929	0.963	0.956
recall	**0.976**	0941	0962	0.949	0.973	0962
F1	**0.974**	0.930	0.958	0939	0968	0.959
AUC scores	**0.996**	0.972	0.971	0978	0.994	0989

It can be seen that in Table 4 the Random Forest algorithm (far left column) is the most superior in terms of accuracy, f1 recall score, and AUC score.

4.2 Based on the URL Text Arrangement

Classification of phishing URLs based on URL text arrangement is performed on the LSTM and GRU deep learning algorithms [20]. Classification results based on URL text are in Table 5.

The results of the deep learning algorithm based on URL characteristics can be seen in Table 3, it is found that the LSTM algorithm has the highest accuracy results. Table 5 shows the GRU algorithm as the algorithm with the best results based on the arrangement of the URL text. Based on these results and research [21], a test was carried out by combining 2 models, namely the GRU model and the LSTM model as a model for classifying.

The GRU-LSTM model in Fig. 3 has 2 input layers. The right layer receives input in the form of a token from the URL and the left layer receives input in the form of 69 feature URLs. The next layer on the right is the embedding layer which converts it into a 2D vector to enter the GRU layer with 256,216 and 128 units. The left layer goes into the Bidirectional LSTM layer with 128 units and a dropout layer. Then enter the Bidirectional LSTM layer with 64 units and the dropout layer and the Dense layer with

Table 5. Comparison of LSTM and GRU

	LSTM	GRU
accuracy	0962	**0978**
Precision	0.961	**0.975**
recall	0962	**0.981**
F1 scores	0962	**0978**
AUC scores	0.992	**0.998**

32 units. The results of the 2 layers are then combined and entered the dense layer. The last layer is a dense layer to determine the classification results.

The optimal libraries used to create deep learning models are TensorFlow and, libraries for creating Fig. 4 show the results of the combination model training process which shows a decrease in validation results under the training results in the model training process. The process of showing validation results under the training results is seen in processes 6 to 9. Therefore, the model that is considered optimal is the model in the 5th training process, where this model is used to test the test data.

The optimal libraries used to create deep learning models are TensorFlow and, libraries for creating Fig. 4 shows the results of the combination model training process which shows a decrease in validation results under the training results in the model training process. The process of showing validation results under the training results is

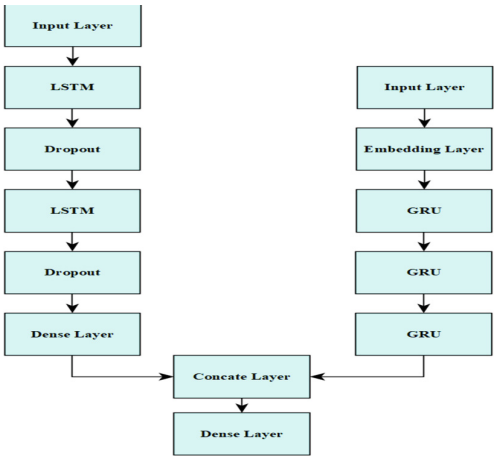

Fig. 3. Architecture of GRU and LSTM.

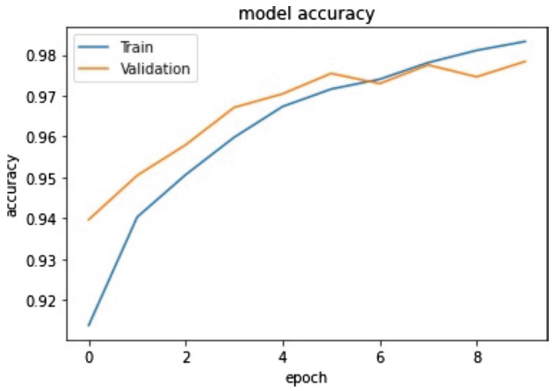

Fig. 4. Model Accuracy.

Table 6. A final comparison of 4 models

	Random Forest	LSTM	GRU	GRU- LSTM
Accuracy	0.974	0.968	0.978	**0.981**
Precision	0.971	0.963	0.975	**0979**
Recall	0.976	0.973	0.981	**0.982**
F1	0.974	0.968	0.978	**0.981**
AUC score	0.996	0.994	0.998	**0.997**

seen in processes 6 to 9. Therefore, the model that is considered optimal is the model in the 5th training process, where this model is used to test the test data.

The LSTM column in Table 6 is the result of a classification based on URL characteristics. The GRU column is the result of a classification based on the arrangement of the text in the URL. The GRU-LSTM column is the result of combining 2 models. From the table above there has been an increase in all aspects assessed in the combination model that has been carried out.

5 Conclusions

Based on the results of the research conducted, several conclusions were obtained to answer the formulation of the problem in this study, deep learning implementation can be used to classify phishing urls with an accuracy of 98.1%. The architecture used is a combination of LSTM and GRU with feature and text-based methods as input.

Based on the results of the benchmarking model, a deep learning model is obtained for phishing url classification using a combination of LSTM and GRU architectures to get higher results than shallow machine learning, namely random forest, respectively with 98.1% and 97.4% accuracy. For auc scores, respectively, namely 99.7% and 99.6%

References

1. N. A. Khan, "Ten Deadly Cyber Security Threats Amid COVID-19 Pandemic," 2020, https://doi.org/10.36227/techrxiv.12278792.v1.
2. H. Abroshan, J. Devos, G. Poels, and E. Laermans, "COVID-19 and Phishing: Effects of Human Emotions, Behavior, and Demographics on the Success of Phishing Attempts during the Pandemic," *IEEE Access*, vol. 9, pp. 121916–121929, 2021, https://doi.org/10.1109/ACCESS.2021.3109091.
3. APWG, "Phishing E-mail Reports and Phishing Site Trends," 2022. [Online]. Available: http://www.apwg.org,
4. F. Mouton, M. M. Malan, L. Leenen, and H. S. Venter, "Social engineering attack framework," in *2014 Information Security for South Africa - Proceedings of the ISSA 2014 Conference*, Nov. 2014. https://doi.org/10.1109/ISSA.2014.6950510.
5. H. Abroshan, J. Devos, G. Poels, and E. Laermans, "Phishing attacks root causes," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 10694 LNCS, pp. 187–202. https://doi.org/10.1007/978-3-319-76687-4_13.

6. R. Rao, T. Vaishnavi, and A. Pais, "CatchPhish: detection of phishing websites by inspecting URLs," *J Ambient Intell Humaniz Comput*, vol. 11, Dec. 2020, https://doi.org/10.1007/s12652-019-01311-4.
7. B. Andriy Burkov, *The Hundred-Page Machine Learning*.
8. E. D. O. Andrade, J. Viterbo, C. N. Vasconcelos, J. Guérin, and F. C. Bernardini, "A model based on LSTM neural networks to identify five different types of malware," in *Procedia Computer Science*, 2019, vol. 159, pp. 182–191. https://doi.org/10.1016/j.procs.2019.09.173.
9. M. Selvakumari, M. Sowjanya, S. Das, and S. Padmavathi, "Retraction: Phishing website detection using machine learning and deep learning techniques," *Journal of Physics: Conference Series*, vol. 1916, no. 1. IOP Publishing Ltd, May 27, 2021. https://doi.org/10.1088/1742-6596/1916/1/012169.
10. 11. Y. Su, "Research on Website Phishing Detection Based on LSTM RNN," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Jun. 2020, vol. 1, pp. 284–288. https://doi.org/10.1109/ITNEC48623.2020.9084799.
11. M. Korkmaz, O. K. Sahingoz, and B. DIri, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis," in *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*, Jul. 2020. https://doi.org/10.1109/ICCCNT49239.2020.9225561.
12. G. Vrban, I. Jr. Fister, and V. Podgorel, "Datasets for phishing websites detection," *Elsevier*, 2020, https://doi.org/10.17632/72ptz43s9v.1.
13. A. Saleem Raja, R. Vinodini, and A. Kavitha, "Lexical features based malicious URL detection using machine learning techniques," *Mater Today Proc*, vol. 47, pp. 163–166, 2021, https://doi.org/10.1016/j.matpr.2021.04.041.
14. W. A. Qader, M. M. Ameen, and B. I. Ahmed, "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges," in *Proceedings of the 5th International Engineering Conference, IEC 2019*, Jun. 2019, pp. 200–204. https://doi.org/10.1109/IEC47844.2019.8950616.
15. S. Edu, "DATA MINING USING A SUPPORT VECTOR MACHINE, DECISION TREE, LOGISTIC REGRESSION AND RANDOM FOREST FOR PNEUMONIA PREDICTION AND CLASSIFICATION." [Online]. Available: https://www.researchgate.net/publication/361179116
16. N. Zhu, C. Zhu, L. Zhou, Y. Zhu, and X. Zhang, "Optimization of the Random Forest Hyperparameters for Power Industrial Control Systems Intrusion Detection Using an Improved Grid Search Algorithm," *Applied Sciences (Switzerland)*, vol. 12, no. 20, Oct. 2022, https://doi.org/10.3390/app122010456.
17. N. K. Manaswi, "RNN and LSTM," in *Deep Learning with Applications Using Python : Chatbots and Face, Object, and Speech Recognition With TensorFlow and Keras*, Berkeley, CA: Apress, 2018, pp. 115–126. https://doi.org/10.1007/978-1-4842-3516-4_9.
18. 19. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, https://doi.org/10.1162/neco.1997.9.8.1735.
19. 20. Y. Kim, M. Chae, N. Cho, H. Gil, and H. Lee, "Machine Learning-Based Prediction Models of Acute Respiratory Failure in Patients with Acute Pesticide Poisoning," *Mathematics*, vol. 10, no. 24, p. 4633, Dec. 2022, https://doi.org/10.3390/math10244633.
20. M. Zulqarnain, R. Ghazali, M. Ghulam Ghouse, and M. Faheem Mushtaq, "Efficient Processing of GRU Based on Word Embedding for Text Classification."
21. 22. A. Ozcan, C. Catal, E. Donmez, and B. Senturk, "A hybrid DNN–LSTM model for detecting phishing URLs," *Neural Comput Appl*, 2021, https://doi.org/10.1007/s00521-021-06401-z.